Notes

Improved estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation

RAMPAL S. ETIENNE

*Community and Conservation Ecology Group, Centre for Ecological and Evolutionary Studies, University of Groningen,
P.O. Box 14, 9750 AA Haren, The Netherlands*

Abstract. The standard neutral model of biodiversity, as introduced by S. P. Hubbell, is currently increasingly used as a null model for the structure of ecological communities. In such applications, estimation of the model parameters is essential. An exact maximum likelihood approach has been developed for data sets consisting of multiple, spatially separated, samples of species abundances. This approach is only computationally tractable when it is assumed that all these samples have the same amount of dispersal limitation. Recently, an approximate approach has been proposed that does not require this assumption. However, this approach cannot estimate the fundamental biodiversity number θ when there are only a few samples or many, very different, samples. In this note, I present a modification of this approximate approach that does not suffer from this shortcoming. I illustrate it with simulated and real data sets.

Key words: Etienne sampling formula; Ewens sampling formula; fundamental biodiversity number; fundamental dispersal number; maximum likelihood.

INTRODUCTION

The now well-known neutral theory of biodiversity (Hubbell 2001) is increasingly being used as a null model for the structure of ecological communities, particularly for the distribution of abundances of the species in a community (Alonso et al. 2006; see Gotelli and McGill 2006 for a discussion of neutral vs. null models). The most widely used model implementation of this theory is a mainland–island model where the local community receives immigrants from the regional community (the metacommunity), which itself is governed by speciation and extinction. Both local and regional abundances vary stochastically, a process termed ecological drift (Hubbell 2001). The model contains two parameters: the fundamental biodiversity parameter (θ) and the fundamental dispersal or immigration parameter (I; Etienne and Alonso 2005). The former parameter summarizes the diversity of the metacommunity community and is a composite of the metacommunity size (J_M) and the speciation rate (ν). The latter parameter measures the amount of dispersal limitation of the local community and may be interpreted as the number of potential immigrants competing with local individuals for vacant sites (resources)

[1] E-mail: r.s.etienne@rug.nl

(Etienne and Olff 2004*a*). When dealing with samples of species abundances, the parameter m has often been used instead of I, where $m = I/(I + J - 1)$ with J the total number of individuals in the species abundance data set. This parameter may be interpreted as the probability that any of the I immigrants will occupy the vacant site in a community of size J.

Applications of the model often involve estimation of these parameters from samples of species abundances because these parameters may provide new and relatively quick insights into the ecology of a system, even if neutrality is violated (Latimer et al. 2005, Jabot et al. 2008). Exact sampling formulae have been developed with which these parameters can be simultaneously estimated using maximum likelihood (Etienne and Olff 2004*b*, Etienne 2005, 2007). These exact formulae are tractable in practice only when there is a single sample, or when there are multiple, spatially separated samples (i.e., from different local communities connected to the same metacommunity) which are assumed to have the same degree of dispersal limitation. That is, only a single value of I can be estimated. Exact and simultaneous estimation of a separate value for I for each local sample is currently only theoretically possible (Etienne 2007). Fortunately, an approximate approach has been proposed that does not require this assumption (Munoz et al. 2007). It is approximate in the sense that it does not

TABLE 1. Estimates of θ and *m* in various scenarios of simulated data sets.

						Maximum likelihood parameter estimation			
						Etienne (2007)			
	Sample sizes and model parameters					$\hat{\theta}_{E2007}$		$\hat{m}$	
Scenario	$\vec{J}$	θ	m_1	m_2	m_3	Mean	CV	Mean	CV
1	(1000, 1000, 1000)	50	0.1	0.1	0.1	50.03	0.1126	0.1010	0.1029
2	(100, 1000, 100)	50	0.1	0.1	0.1	52.53	0.3060	0.1023	0.1680
3	(500, 500, 500)	50	0.1	0.1	0.1	50.39	0.1299	0.1017	0.1143
4	(2000, 2000, 2000)	50	0.1	0.1	0.1	50.09	0.0932	0.1009	0.0911
5	(10 000, 10 000, 10 000)	50	0.1	0.1	0.1	50.03	0.0717	0.1014	0.1053
6†§	(1000, 2000, 3000)	50	0.1	0.1	0.1	51.89	0.0910	0.1698	0.0901
7	(1000, 1000, 1000)	5	0.1	0.1	0.1	5.0434	0.2231	0.1053	0.2513
8	(1000, 1000, 1000)	500	0.1	0.1	0.1	502.16	0.0875	0.1003	0.0539
9	(1000, 1000, 1000)	5	0.01	0.01	0.01	5.0766	0.3082	0.0110	0.4086
10	(1000, 1000, 1000)	50	0.01	0.01	0.01	51.12	0.2055	0.0101	0.1304
11	(1000, 1000, 1000)	500	0.01	0.01	0.01	545.60	0.3642	0.0100	0.1021
12†	(1000, 1000, 1000)	5	0.1	0.2	0.4	5.0988	0.2095	0.1910	0.2538
13†	(1000, 1000, 1000)	50	0.1	0.2	0.4	52.11	0.0981	0.1879	0.1028
14†	(1000, 1000, 1000)	500	0.1	0.2	0.4	535.80	0.0681	0.2011	0.0474
15†	(1000, 1000, 1000)	5	0.01	0.05	0.25	6.1630	0.2498	0.0332	0.3396
16†	(1000, 1000, 1000)	50	0.01	0.05	0.25	68.72	0.1375	0.0413	0.1173
17†	(1000, 1000, 1000)	500	0.01	0.05	0.25	773.20	0.1218	0.0644	0.0492
18‡§	(1000, 2000, 3000)	5	0.1	0.053	0.036	5.0022	0.2239	0.1061	0.2400
19‡§	(1000, 2000, 3000)	50	0.1	0.053	0.036	49.90	0.1056	0.1004	0.0942
20‡§	(1000, 2000, 3000)	500	0.1	0.053	0.036	502.61	0.0765	0.0999	0.0484

Notes: The values reported are the means and coefficients of variation (CV) of the parameter estimates over 1000 simulated data sets. The parameter $\hat{\theta}_{new}$ is calculated exactly (rather than by Monte Carlo sampling) using Eq. 7a. Note that $\hat{\theta}_{E2007}$ cannot be estimated accurately in the scenarios marked with a dagger (†) because the requirements of equal dispersal limitation are not met. In the scenarios marked with "‡," these requirements are met, because sample sizes and dispersal limitation parameters m_i are chosen in such a way that I is the same for all three samples. Note that $\hat{\theta}_{M2007}$ would be infinite in many of the 1000 simulated data sets, and therefore it is not reported. The new two-stage approach proposed in this note has been developed precisely to fix this problem. The values of $\hat{m}$ in the scenarios marked with "§" (scenarios with unequal sample sizes) are for the first sample only. The mean values of $\hat{m}$ for the other samples are as follows: scenario 6, 0.0928, 0.0639; scenario 18, 0.0562, 0.0382; scenario 19, 0.0529, 0.0359; scenario 20, 0.0526, 0.0357.

estimate the parameters (θ and the various I) simultaneously and it needs Monte Carlo samples from the data set to estimate θ. Unfortunately, this approach does not work when there are only a few samples available, or when there are many samples that are very different. Here I show why it does not work, and I present a modification of this approach that solves this problem. I illustrate it with simulation data and real data. A program code is provided in the online material.

IMPROVED TWO-STAGE ESTIMATION

Denote the species abundance data set from which we want to infer the model parameters by the vector $\mathbf{n} = (n_{11}, \ldots, n_{1S}, n_{21}, \ldots, n_{2S}, \ldots, n_{N1}, \ldots, n_{NS})$: Here, N is the number of local samples, S is the total number of different species in the entire data set, and n_{ik} is the abundance of species k in local sample i. The two-stage estimation approach of Munoz et al. (2007) consists of the following two steps.

1. *Munoz et al. estimation of θ.*—Repeat the following procedure M times and denote each iteration with the subscript j. Take a single individual from each local sample, record its species, and estimate $\hat{\theta}_j$ from the resulting species abundance data set of N individuals as

follows. Because each local sample supplies only a single individual, the N individuals thus collected can be regarded as a sample from the metacommunity (they have different immigrating ancestors [Etienne and Olff 2004b, Etienne 2007]). Because the metacommunity is governed by the Ewens sampling formula (Ewens 1972, Hubbell 2001), this sampling formula can be used to estimate θ_j. Denoting the abundance vector of the N individuals in the jth iteration by $\mathbf{N}_j$ the Ewens sampling formula reads for this case as follows:

$$P[\mathbf{N}_j | \theta_j, N] = F_j \frac{\theta_j^{S_j}}{(\theta_j)_N} \qquad (1)$$

where F_j is a prefactor that does not depend on θ, S_j is the number of species in the metacommunity sample, which is usually less than the total number of species S, and $(\theta_j)_N$ is the Pochhammer notation,

$$(\theta_j)_N = \prod_{s=1}^{N} (\theta_j + s - 1) = \frac{\Gamma(\theta_j + N)}{\Gamma(\theta_j)} \qquad (2)$$

where Γ refers to the Gamma function. It can be easily shown (Tavaré and Ewens 1997) that the maximum

Table 1. Extended.

Maximum likelihood parameter estimation							
This paper							
$\hat{\theta}_{new}$		$\hat{m}_1$		$\hat{m}_2$		$\hat{m}_3$	
Mean	CV	Mean	CV	Mean	CV	Mean	CV
52.20	0.2154	0.1107	0.5640	0.1093	0.4965	0.1115	0.6637
65.92	0.5892	0.1050	0.4356	0.1070	0.4386	0.1040	0.3878
53.71	0.2402	0.1051	0.3805	0.1035	0.3467	0.1052	0.3873
51.93	0.1941	0.1140	0.7238	0.1150	0.7095	0.1128	0.6654
51.65	0.1751	0.1167	0.8242	0.1165	0.7850	0.1155	0.7736
51.65	0.1963	0.1113	0.4989	0.1169	0.6512	0.1185	0.7332
5.8355	0.4110	0.1977	1.4560	0.1818	1.4523	0.1951	1.4727
518.39	0.1758	0.1004	0.0974	0.1005	0.0957	0.0998	0.0918
6.2700	0.5299	0.0519	3.4264	0.0439	3.5178	0.0499	3.3659
64.09	0.5534	0.0116	2.2138	0.0107	1.2297	0.0110	1.4487
1842.82	3.0285	0.0101	0.2021	0.0100	0.1898	0.0101	0.1929
5.7954	0.4227	0.1970	1.4360	0.2931	1.1452	0.4106	0.9094
51.89	0.1978	0.1108	0.6054	0.2249	0.6026	0.4251	0.5023
509.29	0.1219	0.1006	0.0887	0.2009	0.0916	0.4026	0.1138
5.8464	0.4368	0.0367	3.4607	0.1332	1.8081	0.3282	1.0592
54.32	0.2907	0.0102	0.2259	0.0570	0.8389	0.2841	0.7169
584.15	0.3611	0.0100	0.1654	0.0504	0.1302	0.2578	0.3154
5.7329	0.4215	0.2109	1.3894	0.1482	1.7670	0.1112	2.0253
52.78	0.2085	0.1074	0.5572	0.0559	0.5131	0.0388	0.6254
515.78	0.1745	0.1003	0.1013	0.0525	0.1004	0.0359	0.1002

likelihood estimate is the solution $\hat{\theta}_j$ (to be obtained numerically to arbitrary precision) of

$$S_j = \hat{\theta}_j[\Psi(\hat{\theta}_j + N) - \Psi(\hat{\theta}_j)] \quad (3)$$

where $\Psi(x)$ is the psi function or digamma function, that is, the derivative of the logarithm of the gamma function (see, e.g., Etienne and Olff 2004a). The final estimate of $\hat{\theta}$ is then simply calculated as the average of the M values of $\hat{\theta}_j$ ($j = 1 \ldots M$):

$$\hat{\theta}_{M2007} = \frac{1}{M}\sum_{j=1}^{M}\hat{\theta}_j. \quad (4)$$

Eq. 3 is not only the equation for the maximum likelihood estimator; it is also the equation for the estimator based on the method of moments, that is, it equates the observed number of species S_j to the expected number of species in a (metacommunity) sample of size N. Thus, there is no difference between these two estimation methods in this case.

2. *Munoz et al. estimation of I_i.*—The fundamental dispersal number I_i for each local sample i can now be easily estimated with maximum likelihood by applying my single-sample sampling formula (Etienne 2005) to each local sample i keeping θ fixed to $\hat{\theta}_{M2007}$, because given θ, the local samples are independent and can be treated as single samples.

1′. *Improved estimation of θ.*—The approach outlined above works as long as $S_j < N$, because as soon as $S_j = N$, the estimate of θ becomes infinite, and any average no longer makes sense. To circumvent this problem, I propose not to estimate each $\hat{\theta}_j$, but to write down the likelihood for all M iterations,

$$P[\mathbf{N}_1, \ldots, \mathbf{N}_M | \theta, N] = \prod_{j=1}^{M} F_j \frac{\theta^{S_j}}{(\theta)_N} \quad (5)$$

and then use maximum likelihood to estimate θ. By setting the derivative of this likelihood with respect to θ to zero, one can show that the maximum likelihood estimator is given by the solution of

$$\frac{1}{M}\sum_{j=1}^{M}S_j = \hat{\theta}_{new}[\Psi(\hat{\theta}_{new} + N) - \Psi(\hat{\theta}_{new})]. \quad (6)$$

Again, this solution can be obtained numerically to arbitrary precision. Thus, instead of averaging the estimates $\hat{\theta}_j$, one must average the number of species S_j. If M is large enough, this average is smaller than N and finite values of $\hat{\theta}_{new}$ will be obtained. Only if the local samples have no species in common at all, we find an estimate of $\hat{\theta}_{new} = \infty$, but this is indeed the best estimate in this case. Note again that Eq. 6 also gives the estimator based on the method of moments: the observed number of species averaged over M simulated samples of size N is equated to the number of species in a metacommunity sample of size N predicted by the model. This interpretation gives way to an exact (rather than by Monte Carlo sampling) computation of $\hat{\theta}_{new}$, because the mean number of species in a sample of size N where each individual is taken from a different local community i, can in principle be computed exactly:

$$\langle S(\mathbf{k}) \rangle = \sum_{\{\mathbf{k}\}} \prod_{i=1}^{N} \frac{n_{ik_i}}{J_i} S(\mathbf{k}) \quad (7a)$$

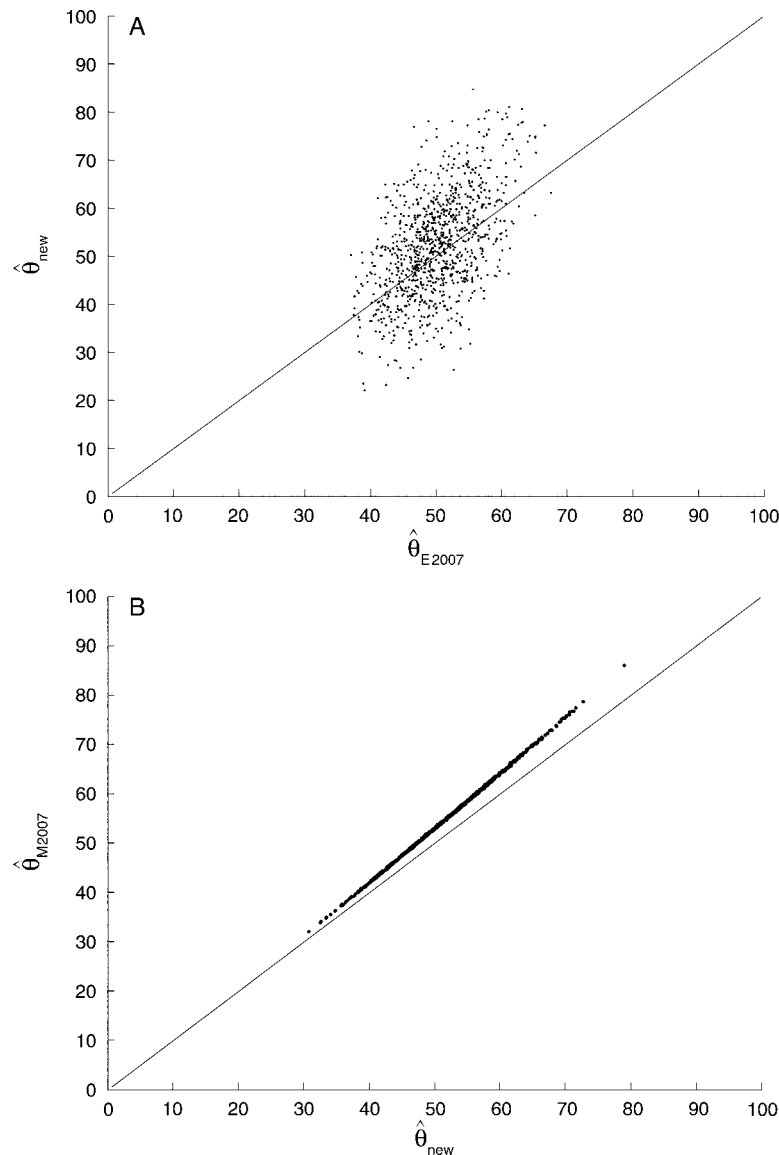where $S(\mathbf{k})$ is the number of species in a sample with vector

Fig. 1. Parameter estimates for the old two-stage approach of Munoz et al. (2007) and the new two-stage approach proposed here. Panel A shows $\hat{\theta}_{new}$ and $\hat{\theta}_{E2007}$ for 1000 simulated data sets of three samples, generated with $\theta = 50$, $J_i = 1000$, and $m_i = 0.1$ ($i = 1, 2, 3$). Panel B shows $\hat{\theta}_{new}$ and $\hat{\theta}_{M2007}$ for 1000 simulated data sets of 50 samples, generated with $\theta = 50$, $J_i = 400$, and $m_i = 0.1$ ($i = 1, \ldots, 50$). The parameter θ is the fundamental biodiversity parameter; J_i and m_i are the size and dispersal limitation parameter of sample i, respectively. The straight line is plotted to guide the eye.

$\mathbf{k} = (k_1, \ldots, k_N)$ with k_i the species label of the individual from site i; the sum is over all possible vectors $\mathbf{k}$. Thus, $\hat{\theta}_{new}$ is the solution of

$$\langle S(\mathbf{k}) \rangle = \hat{\theta}_{new}[\Psi(\hat{\theta}_{new} + N) - \Psi(\hat{\theta}_{new})]. \qquad (7b)$$

In practice, the number of different vectors $\mathbf{k}$ increases very fast with N, so this exact method is only feasible for small N. This is exactly the situation in which the Monte Carlo method requires many iterations to yield an accurate estimate of $\hat{\theta}_{new}$.

2′. *Estimation of I_i.*—The estimation of I_i for each sample i proceeds as before where one fixes θ at $\hat{\theta}_{new}$. In the Supplement, I have provided a code to perform the estimation of θ and I_i.

Results

I illustrate the new approach with simulated data and real data and compare them to the one-stage estimation approach of Etienne (2007) and to the two-stage approach of Munoz et al. (2007). I will refer to each of these three approaches as the new two-stage approach, the one-stage

TABLE 2. Neutral parameter estimates for samples from three local tree communities (Sherman, BCI, and Cocoli) in the Panama Canal Zone using the one-stage approach, which assumes that all I_i are identical, and the new two-stage approach presented in this paper, with exact calculation of $\hat{\theta}_{new}$.

| | Sample sizes and species richness | | Maximum likelihood parameter estimation | | | | | |
| | | | Etienne (2007) | | This paper | | | |
Data set	$\vec{J}$	$\vec{S}$	$\hat{\theta}_{E2007}$	$\hat{I}$	$\hat{\theta}_{new}$	$\hat{I}_{Sherman}$	$\hat{I}_{BCI}$	$\hat{I}_{Cocoli}$
Sherman + BCI + Cocoli	(2860, 21 457, 1079)	(125, 225, 99)	259.3	44.24	341.7	33.87	53.71	30.79
Sherman + BCI$_1$ + Cocoli	(2860, 2359, 1079)	(125, 152, 99)	270.5	39.18	247.7	37.09	53.66	32.61
Sherman + BCI$_2$ + Cocoli	(2860, 2151, 1079)	(125, 150, 99)	273.9	39.21	232.6	37.89	55.43	33.06
Sherman + BCI$_3$ + Cocoli	(2860, 2076, 1079)	(125, 162, 99)	280.0	41.18	322.6	34.35	57.49	31.07
Sherman + BCI$_4$ + Cocoli	(2860, 2027, 1079)	(125, 171, 99)	282.2	42.63	359.8	33.46	61.38	30.55
Sherman + BCI$_5$ + Cocoli	(2860, 2000, 1079)	(125, 166, 99)	290.8	41.71	348.5	33.71	59.40	30.70
Sherman + BCI$_6$ + Cocoli	(2860, 2050, 1079)	(125, 153, 99)	297.3	39.13	399.5	32.71	49.31	30.12
Sherman + BCI$_7$ + Cocoli	(2860, 2364, 1079)	(125, 147, 99)	298.6	37.27	448.4	31.98	42.26	29.69
Sherman + BCI$_8$ + Cocoli	(2860, 2225, 1079)	(125, 138, 99)	296.5	36.32	427.1	32.28	39.83	29.86
Sherman + BCI$_9$ + Cocoli	(2860, 2076, 1079)	(125, 145, 99)	300.4	37.65	427.7	32.27	44.01	29.86
Sherman + BCI$_{10}$ + Cocoli	(2860, 2129, 1079)	(125, 157, 99)	271.5	40.47	369.3	33.27	51.94	30.44

Note: The different BCI$_i$ denote different subsamples from the full BCI (Barro Colorado Island) 50-ha plot.

approach, and the old two-stage approach, respectively. I mainly focus on the estimation of θ, because this is where the new two-stage approach developed here differs from the old two-stage approach, and because the values of m_i (or I_i) depend on the estimate for θ. For the simulated data, I selected two types of data that differ in the number of local samples. In order to show that the new two-stage approach presented here works well when only a few local samples are available, in which case the old two-stage approach does not work at all, I first generated 1000 simulated data sets of three local samples each using the urn method (based on coalescence) explained in Etienne (2007). I did this for various scenarios of sample sizes and parameter values (see Table 1). For each data set, I calculated $\hat{\theta}_{new}$ with the exact method of Eq. 7a, which can be done within reasonable time because $N = 3$ is small. For comparison, I also calculated the exact estimate $\hat{\theta}_{E2007}$ using the full sampling formula of the one-stage approach.

One observes that the averaged estimates for the simulated data sets produce a biased value of θ, but some bias can be expected because it also occurs for the Ewens sampling formula (Tavaré and Ewens 1997). This bias decreases with sample size as scenarios 1–5 show; a sample size of 500 each already gives averaged values of $\hat{\theta}_{new}$ and $\hat{\theta}_{E2007}$ with little bias. However, the coefficient of variation in $\hat{\theta}_{new}$ is substantially larger than that in $\hat{\theta}_{E2007}$. This means that $\hat{\theta}_{new}$ is less reliable than $\hat{\theta}_{E2007}$. Fig. 1A shows the relationship between $\hat{\theta}_{new}$ and $\hat{\theta}_{E2007}$ for the 1000 simulated data sets in scenario 1. The two parameter estimates are clearly correlated, but $\hat{\theta}_{new}$ tends to be larger (smaller) than $\hat{\theta}_{E2007}$ when both are larger (smaller) than the true value θ. The estimates for $\hat{\theta}_{new}$ become worse when diversity is lower (lower θ and lower m). In contrast, $\hat{\theta}_{E2007}$ seems rather insensitive to the diversity of the community, as far as the bias concerned; however, the coefficient of variation increases when diversity decreases. The estimates for the dispersal limitation parameters, $\hat{m}_i$, contain little bias in general;

only when θ is low, the bias is substantial. Again, for low diversity the coefficient of variation of $\hat{m}_i$ is high.

In sum, when the assumption of equal dispersal limitation is met (i.e., same I_i, even if m_i values are different, as in scenarios 18–20 in Table 1), the one-stage approach gives more reliable parameter estimates than the approach presented here, but the latter approach nevertheless performs fairly well as long as diversity is not too low and sample sizes are sufficiently large. When this assumption is not met (the exact situation for which the approach presented here was developed), the new two-stage approach gives better estimates than the one-stage approach.

To be able to compare the estimates based on the new two-stage approach presented here with the old two-stage approach, I generated 1000 simulated data sets of 50 local samples, each local sample containing 400 individuals, using θ = 50 and $m = 0.1$. I estimated both $\hat{\theta}_{M2007}$ and $\hat{\theta}_{new}$ for each of these 1000 simulated data sets, using $M = 10\,000$ iterations. Fig. 1B shows the correlation between the estimates for these 1000 simulated data sets. The mean value of $\hat{\theta}_{new}$ (51.55) is closer to the true value (50) than is $\hat{\theta}_{M2007}$ (54.81), which is therefore more biased. Also the variance of $\hat{\theta}_{new}$ (55.21) is lower than that of $\hat{\theta}_{M2007}$ (68.21). The two estimates are highly correlated as may be expected.

For the real data, I used two examples. The first consists of the three tropical tree communities in the Panama Canal Zone (Condit et al. 2002), which I also used in Etienne (2007). The three plots, Sherman (5.96 ha of which 5 ha is in the data file), Barro Colorado Island (50 ha), and Cocoli (4 ha) lie along a precipitation gradient (3030 mm/yr, 2616 mm/yr, and 1950 mm/yr, respectively [Condit et al. 2004]). Table 2 shows the results of the parameter estimation, for these three plots and for various local samples from the large BCI plot (see Etienne 2007). The estimates for the one-stage approach and the new two-stage approach proposed in this note are very different. This is partly because the

bias and variance are higher for the new two-stage approach (see Table 1), but mostly because the former approach assumes that I is the same for all samples, whereas the latter approach shows that this assumption is not justified: BCI clearly has a higher fundamental immigration number than Sherman and Cocoli. Because I in fact measures recruitment limitation which involves dispersal and establishment (Jabot et al. 2008), this may be explained by the central location of BCI: because of this central location, it receives dispersers from all directions, and it has intermediate precipitation, allowing establishment of species adapted to both wetter and drier circumstances. One needs to realize, however, that it is also possible that the model assumption of a single metacommunity for the three local communities is incorrect. Although studying other models of community structure is beyond the scope of this note, this is a interesting direction for future research.

The second set of real data is the Western Ghat forest data set of Munoz et al. (2007) that consists of 50 local samples with sizes varying from 104 to 876 (average is 391.1) individuals, totaling 19 555 individuals. The number of species per local sample ranges from 19 to 71 and averages 45.24; the total number of species across all samples is 304. With $M = 100\,000$ iterations, I found $\hat{\theta}_{new} = 51.5$. Furthermore, I recalculated $\hat{\theta}_{M2007}$ and found $\hat{\theta}_{M2007} = 54.7$. Compare this with the value $\hat{\theta}_{M2007} = 62.3$ reported in Munoz et al. (2007), which was based on $M = 1000$ iterations. For this data set, the one-stage approach yields $\hat{\theta}_{E2007} = 85.2$, which is much higher than the other estimates (due to the fact that it assumes equal dispersal limitation for all local communities).

CONCLUSION

In this note I have improved the approximate two-stage approach of Munoz et al. (2007) to estimate the parameters of the neutral model (θ and m [or I]) from species abundances in multiple spatially separated local samples. The improvement is that it can also be applied to a small number of local samples, or to many local samples that are very different. In these cases, it is likely that a single draw from each local sample contains only singletons resulting in an infinite $\hat{\theta}_{M2007}$, but a finite $\hat{\theta}_{new}$, the estimate produced by the new two-stage approach. For an accurate estimation of $\hat{\theta}_{new}$, the number of iterations M needed to obtain the expected number of species when drawing a single individual from each local sample must be very large. Fortunately, when the number of local samples N is small, an exact calculation (i.e., without iterations) of $\hat{\theta}_{new}$ is possible. The estimate for $\hat{\theta}_{new}$ has little bias when the diversity is not too low (θ

~ 50, $m \sim 0.1$) and as long as sample sizes are sufficiently large ($J > 1000$). The larger the sample size, the smaller the bias. However, the coefficient of variation is always quite large. I therefore suggest caution in using the two-stage approach when the sizes of the local samples are very small or when diversity is low. When the I_i values are likely to be very similar across samples, it is better to use the one-stage approach of Etienne (2007).

LITERATURE CITED

Alonso, D., R. S. Etienne, and A. J. McKane. 2006. The merits of neutral theory. Trends in Ecology and Evolution 21:451–457.

Condit, R., S. Aguilar, A. Hernandes, R. Perez, S. Lao, G. Angehr, S. Hubbell, and R. Foster. 2004. Tropical forest dynamics across a rainfall gradient and the impact of an El Niño dry season. Journal of Tropical Ecology 20:51–72.

Condit, R., et al. 2002. Beta-diversity in tropical forest trees. Science 295:666–669.

Etienne, R. S. 2005. A new sampling formula for neutral biodiversity. Ecology Letters 8:253–260.

Etienne, R. S. 2007. A neutral sampling formula for multiple samples and an "exact" test of neutrality. Ecology Letters 10:608–618.

Etienne, R. S., and D. Alonso. 2005. A dispersal-limited sampling theory for species and alleles. Ecology Letters 8:1147–1156.

Etienne, R. S., and H. Olff. 2004a. How dispersal limitation shapes species: body size distributions in local communities. American Naturalist 163:69–83.

Etienne, R. S., and H. Olff. 2004b. A novel genealogical approach to neutral biodiversity theory. Ecology Letters 7:170–175.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87–112.

Gotelli, N. J., and B. J. McGill. 2006. Null versus neutral models: what's the difference? Ecography 29:793–800.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.

Jabot, F., J. Chave, and R. S. Etienne. 2008. Reconciling neutral community models and environmental filtering: theory and an empirical test. Oikos 117:1308–1320.

Latimer, A. M., J. A. Silander, and R. M. Cowling. 2005. Neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot. Science 309:1722–1725.

Munoz, F., P. Couteron, B. R. Ramesh, and R. S. Etienne. 2007. Inferring parameters of neutral communities: from one single large to several small samples. Ecology 88:2482–2488.

Tavaré, S., and W. J. Ewens. 1997. Multivariate Ewens distribution. Pages 232–246 in N. Johnson, S. Kotz, and N. Balakrishnan, editors. Discrete multivariate distributions. Wiley, New York, New York, USA.

SUPPLEMENT

Program code for maximum likelihood estimation of neutral model parameters for multiple samples of species abundances using the two-stage approach described in the paper (*Ecological Archives* E090-058-S1).